Linear Regression from scratch using CPP

Project report

OOPS Project Evaluation

November 2020

by

V SATYA PAVAN KALYAN

(2019BCS-069)

under the supervision of

Dr. VINAL PATEL

विश्वजीवनामृतं ज्ञानम्

ABV-INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT

GWALIOR-474 015

CANDIDATE'S DECLARATION

I hereby certify that I have properly checked and verified all the items as prescribed in the check-list and ensure that my thesis/report is in proper format as specified in the guideline for thesis preparation.

I also declare that the work containing in this report is my own work. I, understand that plagiarism is defined as any one or combination of the following:

1. To steal and pass off (the ideas or words of another) as one's own
2. To use (another's production) without crediting the source
3. To commit literary theft
4. To present as new and original an idea or product derived from an existing source.

I understand that plagiarism involves an intentional act by the plagiarist of using someone else‘s work/ideas completely/partially and claiming authorship/originality of the work/ideas. Verbatim copy as well as close resemblance to some else‘s work constitute plagiarism.

I have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programs, experiments, results, websites, that are not my original contribution. I have used quotation marks to identify verbatim sentences and given credit to the original authors/sources.

I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report/dissertation/thesis are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable. My faculty supervisor(s) will not be responsible for the same.

Signature:

Name: VEMULA SATYA PAVAN KALYAN
Roll No.: 2019BCS-069
Date: 14/11/2020

ABSTRACT

Even with huge hype on machine learning in the industry and developer community, there hasn't been any standard library for machine learning in C++ language. In order to address the lack of exploration of machine learning implementation in C++ language, the document provides an insight on building a basic machine learning module in C++ language which comprises of a linear regression methods. The module was built using the core concepts of C++ such as object oriented programming. Linear regression is one of the most common guesstimate performed by anyone, it's the basis for regression problems and many analyses.

Contents

1 Introduction

Linear regression is used for finding linear relationship between the target variable and one or more predictor variables. Namely there are two types of linear regression, the former being "Simple linear regression" and the latter being "Multiple linear regression".

The main idea behind linear regression can be condensed as "the process of finding a line that best fits the given data". A best fit line can described as the line for which the total prediction error (for all data points) is zero or as small as possible. Error is deviation of the predicted value from actual value, this is closely related to the distance between the actual data point to the regression line.

We will implement the **Simple Linear Regression model**. Simple linear regression concerns two-dimensional sample points with one independent variable and one dependent variable and finds a linear function that predicts the dependent variable values as a function of the independent variable.

In layman terms, we have a dependent variable and an independent variable. We try to express the dependent variable in terms of the independent variable by expressing it in an line equation

$$Y = mX + C$$

where **X** is the explanatory variable and **Y** is the dependent variable

Different approaches to Linear Regression:

1. Solving model parameters
2. Using optimization algorithm

The methods adopted

1. Closed-form, which is an example of solving model parameters
2. Gradient descent, which is an example of Using optimization algorithm

We use R^2 method for model evaluation, where $R^2 = 1 - \sum(Y_p - Y)^2 / \sum(Y - \bar{Y})^2$

2 Basic Implementation

CSV Read class
It contains vector pair to take an n-columned data from the csv and store it and also string to save the file name and vectors x,y with which we are computing
Functions

- read – function to read the input csv file and save the data into vectors
- split – To split the given dataset into training and test data set
- data normalization – normalizing dataset to fit in the range of C++ data types

LinearRegression Class
It stores the parameters of the predicted line like slope,intercept,mean of x,mean of y and it contains member functions
Functions:

- calculate – to find the parameters of the predicted line
- show – display function to cross verify things
- predicted – using one the variables to find the predicted variable
- predict gradient – find the test dataset values using the slope and intercept found by gradient descent

Accuracy class
This class stores the correlation, R^2 for closed-form equation ,R^2 for gradient descent and overall accuracy of the model . functions

- correlation
- rsquare for formula
- rsquare for gradient

3 Implementation

3.1 Data Wrangling

The target here is to read a CSV file and saved the data in vectors,normalize the data and split the dataset into 2 parts one for training and the other for testing so that we can perform mathematical operations on it.So, here we are making vector pair of string and vector array so that we can expand this function multiple linear regressions with n variables.

Our approach:

- using cpp 's file handling functions opening and accessing the csv file contents
- using getline function to read line by line
- storing each line in a string-stream **s**
- the first line consists of the column names which are added to the vector pairs with column name and an empty integer vector
- the following line consists of integer which are inputed into a string stream
- the string stream numbers are saved into a temporary variable and pushed into respective vectors
- dataset is normalized based on the normalizing factor that we provide and its different for different dataset
- dataset is then split into 2 for training and testing respectively

3.2 Linear Regression class

The main target of this class is to find the line equation of the predicted line i.e.., to find slope and intercept of the predicted line
Our Approach:
1) Closed form equation

- Here we try the find the predicted line equation by direct calculations instead of hit and trial method also called as model calculations
- slope=$\sum(X-\bar{X})(Y-\bar{Y})/\sum(X-\bar{X})^2$
- from Y=MX+C we derive intercept(C) $= \bar{Y} - M * \bar{X}$ value of the predicted line equation

2)Gradient descent

- The best-fit line of the model is found out based on the given training data

- First the function takes random values of slope and intercept and slowly function is trying to reduce the cost based on the cost function we used in this case $(Y_p - Y)^2$ and we try to reduce the cost by iteratively updating the slope and intercept of the line

- The gradient decent method uses estimated error and fixed learning rate to find a suitable δm and δc to reach the error minima.

3.3 Accuracy

R^2

- R^2 evaluates the scatter of the data points around the fitted regression line predicted line
- $R^2 = 1 - \sum(Y_p - Y)^2 / \sum(Y - \bar{Y})^2$
- Generally it ranges [0,1]
- And higher the value of R^2 better is the fit to the data

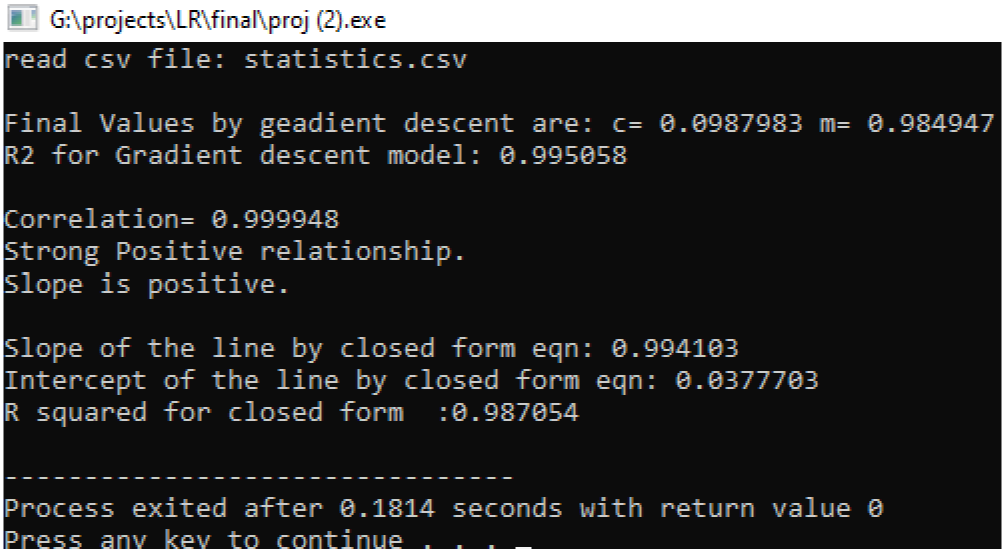
Correlation

- It is helpful to find whether two parameters are linearly dependent on one another or not
- It also helps to find the slope positive or negative
- correlation $= \sum(X - \bar{X})(Y - \bar{Y}) / \sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}$
- If correlation is between [-1,- 0.5] the parameters are strongly linearly related with negative slope
- If correlation is between [-0.5, 0.5] the parameters are not linearly related
- If correlation is between [0.5,1] the parameters are strongly linearly related with positive slope

4 Proposed Approach

The main idea is to read the csv file and divide the given data into training and test data set.To fit a straight line to the given training dataset to predict the values on the test dataset using closed form equation and gradient descent approach individually and find their respective accuracy using R^2 cost function

5 Results



```
read csv file: statistics.csv

Final Values by geadient descent are: c= 0.0987983 m= 0.984947
R2 for Gradient descent model: 0.995058

Correlation= 0.999948
Strong Positive relationship.
Slope is positive.

Slope of the line by closed form eqn: 0.994103
Intercept of the line by closed form eqn: 0.0377703
R squared for closed form  :0.987054

--------------------------------
Process exited after 0.1814 seconds with return value 0
Press any key to continue . . .
```

6 Conclusions

Gradient descent and closed form equation have achieved similar slope and intercept and the data is good for linear regression and we have achieved and high accuracy has been achieved

In C++ we have problems with data types and the limitation of its size which made ML enthusiast to adopt python recently many new libraries like boost have been popular to solve this problem we have used the inbuilt data types and adopted data normalization to counter this problem which shows even though C++ has its limitations its possible to implement ML libraries just it takes more effort on the developer than on the user

7 References

https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86